
Educational Research and Evaluation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nere20

# Benchmarking the performance of Bhutanese students with the performance of the students from the OECD's PISA countries

Gembo Tshering [a] & Vaughan Prain [a]
[a] School of Education, Faculty of Education, La Trobe University, Bendigo, Australia

Available online: 13 Oct 2011

To cite this article: Gembo Tshering & Vaughan Prain (2011): Benchmarking the performance of Bhutanese students with the performance of the students from the OECD's PISA countries, Educational Research and Evaluation, 17:4, 263-281

To link to this article: http://dx.doi.org/10.1080/13803611.2011.621763

*Educational Research and Evaluation*
Vol. 17, No. 4, August 2011, 263–281

Routledge
Taylor & Francis Group

# Benchmarking the performance of Bhutanese students with the performance of the students from the OECD's PISA countries

Gembo Tshering* and Vaughan Prain

*School of Education, Faculty of Education, La Trobe University, Bendigo, Australia*

(*Received 25 February 2011; final version received 25 February 2011*)

Setting international benchmarks for education systems of the Organisation for Economic Co-operation and Development (OECD) countries is one of the goals of the OECD's Programme for International Student Assessment (PISA). However, some countries are not able to participate in PISA, despite their desire to set international benchmarks for their education systems. This article presents a method of setting international benchmarks for a country's school education system, without necessarily participating in PISA, by designing a test using the test items released by PISA for public consumption. The method has been implemented in a study that involved 1,500 Grade 10 students across 60 schools in Bhutan. The students were administered a mathematics test constructed from the PISA Mathematical Literacy test items. The study showed that the performance of Bhutanese students was comparable with the performance of the students from the countries that participated in PISA 2003 and that Bhutan could learn from both high- and low-performing school education systems of those countries.

**Keywords:** benchmarking; PISA; test validity; linking tests; item calibration; alignment between test and curriculum standards; Bhutanese education system; Item Response Theory; plausible values



## Introduction

One of the primary goals of the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) has been to enable its participant countries to compare and learn from each other's strengths and weaknesses in preparing their school children for competent participation in the global economy (OECD, 2004a, 2007). Similar views are emphatically expressed by Creemers (2006) and Kyriakides (2006). Based on its primary goal, PISA has been designed to assess a range of cognitive and noncognitive characteristics of students together with diverse characteristics of teachers and schools that are widely understood to influence school education systems. However, countries with small student enrolment number in the PISA age cohort are not able to participate in PISA because such countries do not have the required number of participants needed to achieve sampling accuracy and precision to get valid estimates of student achievement. For instance, PISA prescribes

*Corresponding author. Email: gtshering@students.latrobe.edu.au

ISSN 1380-3611 print/ISSN 1744-4187 online

http://dx.doi.org/10.1080/13803611.2011.621763
http://www.tandfonline.com

a "minimum of 150 schools" for a sample survey, or otherwise it has to be a census survey (OECD, 2005c, p. 48). Obviously, the latter will involve more resources and time than the former (Ross, 1992), making it more difficult for small countries with limited resources to participate in PISA. Consequently, small countries are deprived of the rich and varied information offered by PISA to its participating countries, making them unable to benchmark their school education systems with the education systems of the countries that participate in PISA.

However, PISA releases some of its test items, with their psychometric properties (e.g., item difficulty, item *p* value), for educational use (see OECD, 2009b) by interested individuals. It is possible to construct a new test with the PISA-released items and link the new test with the OECD's PISA by using *Item Response Theory* (see Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Van der Linden & Hambleton, 1997). Methods of linking tests are widely reported in literature (Johnson & Phillips, 1998; Kolen & Brennan, 2004; OECD, 2009a; Phillips, 2007). It is, however, important to ensure the relevance of a test constructed from the PISA-released items to the school curriculum that the prospective test candidates have followed. A test has to be valid. The American Educational Research Association, (AERA), the American Psychological Association, (APA), and the National Council on Measurement in Education (NCME) (1999) describe validity as the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). This definition emphasises validity as a unitary concept in terms of evidence and extends over construct validity (Aiken & Growth-Marnat, 2006; Gronlund & Waugh, 2009; Linn, 2002), content validity (Aiken & Growth-Marnat, 2006; Gronlund & Waugh, 2009; Linn, 2002), and criterion validity (Aiken & Growth-Marnat, 2006; Gronlund & Waugh, 2009; Linn, 2002). AERA, APA, and NCME (1999) identified the following sources of validity evidence: contents of test, response processes/patterns observed in test, internal structure of test, relation of test to external criterion, and consequences of test. Relating these sources to PISA, similar evidences have been convincingly described in the PISA's assessment frameworks (OECD, 1999, 2004b, 2006), reports (OECD, 2000, 2004a, 2007), and technical reports (OECD, 2002, 2005c, 2009a). However, a validity issue often raised with PISA is its scope to align with school curricula used by participant schools (McGaw, 2008; Nardi, 2008; Prais, 2003; Wagemaker, 2008).

Webb (1997, 1999) presented a convincing discussion on the methods of assessing alignment of tests and curriculum standards, which are used in this article for the same purpose. Aligning the PISA-released items to school curriculum has the flexibility of directly relating the test result to school curriculum standards, which is imperative for providing feedback to schools and other stakeholders (parents, policy-makers). By developing a test with a set of PISA-released test items, by aligning the test to school curriculum standards, and by linking the test to the OECD's PISA results, it is possible to benchmark a nation's education system with the education systems of the countries that participated in PISA. Such an approach has the potential to help the countries without enough number of students in the PISA age cohort in benchmarking their education systems with the countries that participated in PISA. This approach has been tried out in Bhutan.

The aim of this article is to present an overview of a method used in benchmarking the performance of Grade 10 Bhutanese students with the performance of the students of the countries that participated in PISA by using a valid test constructed with a set of the PISA-released items.

## Methods

### *Participants*

Sixty schools and 1,500 (boys = 771, girls = 725) Grade 10 Bhutanese students were sampled from 71 schools and 9,213 students by using two-stage cluster sampling design with equal probability of selection. In the first stage, 60 schools were selected with the selection probability proportional to their sizes. In the second stage, a cluster of 25 students was selected from each of the 60 schools by simple random sampling. The number of valid response papers collected was 1,496.

### *Test validity*

A 2-hr mathematics test was developed by using 42 PISA-released items for a PhD research. Among the main purposes that the test was designed for, setting international benchmarks for Grade 10 Bhutanese students' mathematical knowledge and skills forms the theme of this article. The items were part of a pool of items used by the OECD in the PISA 2003 cycle to assess 15-year-old students' mathematical knowledge and skills described in *The PISA 2003 Assessment Framework* (OECD, 2004b). To ensure that the mathematics test measured the mathematical knowledge and skills that Grade 10 Bhutanese students were expected to learn, a content validation study of the test was conducted.

The following two approaches were used in the content validation study: (a) Bhutan's Grade 10 mathematics curriculum standards (BMCS) were compared to the PISA mathematical literacy domains (PML), and (b) the alignment of the mathematics test items to BMCS was evaluated. The findings from the comparison are expected to indicate similarities and differences between PML and BMCS, while the findings from the alignment study are expected to confirm alignment or non-alignment of the mathematics test items to BMCS.

BMCS and PML were compared in terms of their objectives and content domains by using a method that Osta (2007) would have termed as text analysis. First, the objectives of BMCS were compared to the objectives of PML by identifying similarities and differences in their keywords. Second, the domains of BMCS were compared to the domains of PML by noting the similarities and differences in the mathematical knowledge and skills emphasized in the two domains.

The presence of similarities in the objectives and the domains of BMCS and PML is not sufficient to conclude that a set of test items designed for assessing the latter can assess the former. It is imperative that the mathematics test aligned well with BMCS as much as the test items aligned with PML. Webb (1997, 1999, 2006) effectively demonstrated the use of the following four criteria to assess the alignment of test items to curriculum standards: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. Webb (1999) defined each of these four criteria as follows:

> The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.... Depth-of-knowledge consistency between standards and assessment indicates alignment, if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.... The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students

> need in order to correctly answer the assessment items/activities . . . . The balance-of-representation criterion is used to indicate the extent to which items are evenly distributed across objectives. (pp. 7–8)

Webb (1999) formulated a set of guidelines for evaluating whether or not a test and a set of curriculum standards passed the four alignment criteria. First, there should be at least six items on a test measuring contents of a single curriculum standard for there to be a categorical concurrence between the standard and the test. Second, at least 50% of the items on a test measuring the standard should have their depth-of-knowledge levels at or above the depth-of-knowledge levels of the corresponding objectives of the standard for the depth-of-knowledge consistency between the standard and the test to exist. Third, at least 50% of the objectives of a curriculum standard should have at least one item each on the test for the range-of-knowledge criterion to pass. Fourth, an index of 0.70 or higher, based on the difference in the proportion of objectives and the proportion of hits assigned to the objective, is required for the balance of representation criterion to pass. To evaluate the overall alignment of a test to a set of curriculum standards, Webb (1999) calculated alignment percentages for alignment criteria across standards and graded the alignment into fully aligned (100%), highly aligned (70% to 99%), partially aligned (50% to 69%), and poorly aligned (less than 50%). These guidelines were used in evaluating the alignment of the mathematics test items to BMCS.



### *Participants in the alignment study*

Five mathematics teachers, who have been teaching Grade 10 mathematics in five schools in Bhutan for over a range of 5 to 10 years, were recruited to participate in the study. Each one of the five teachers was provided with a copy of BMCS a week before the study, with the instruction to review them. The objective of the advance distribution of BMCS was to enable the teachers to recapitulate their experience with the curriculum standards and prepare them for the alignment study.

### *Familiarising teachers with test item competency clusters*

On the day of the alignment study, the teachers were presented with the depth-of-knowledge levels used in the mathematics test items. PISA (OECD, 1999, 2004b, 2006) has used the following three broad mathematical competency clusters to evaluate the depth-of-knowledge levels: *reproduction cluster, connections cluster*, and *reflection cluster*. The reproduction cluster is the lowest level of knowledge, requiring students to demonstrate their ability to recall standard representation and definitions, perform routine computations, apply routine mathematical procedures, and solve routine mathematical problems. The connections cluster is the middle level of knowledge, requiring students to display their ability to apply mathematical models, carry out standard problem-solving and interpretation, and apply multiple well-defined methods. The highest level of knowledge is the reflection cluster, requiring students to do complex problem-solving and posing, demonstrate reflection and insight, use original mathematical approaches, apply multiple complex methods, and generalize.

Each one of the five teachers was instructed to study the competency clusters and add their own additional clusters or descriptions under any clusters. The study helped teachers to understand the competency clusters and enabled the researcher to

find out if the teachers agreed or disagreed with the clusters. The teachers were then instructed to discuss their understanding of the competency clusters in groups. The group discussion further clarified the meanings of the clusters and enabled the teachers to reach consensus on the meanings of the clusters. It is worth noting here that none of the five teachers proposed any additional clusters or description within any cluster, indicating the adequacy of the clusters in covering a range of mathematical knowledge and skills expected of Grade 10 Bhutanese students.

*Assigning depth-of-knowledge level to BMCS objectives*

The teachers were instructed to assign individually a depth-of-knowledge level to each objective of each standard of BMCS. Where an objective involved two or more depth-of-knowledge levels, the teachers were instructed to assign the higher level because the lower level is a prerequisite for the higher level. The teachers were instructed, then, to discuss the depth-of-knowledge levels that they assigned to each objective to each standard, with the goal of reaching consensus. The depth-of-knowledge levels assigned to each objective to each standard by the teachers were coded on a coding matrix that was prepared prior to the day of the alignment study. BMCS had a hierarchical structure, with objectives forming the first level, goals the second level, and standards the third level. BMCS had six standards, 29 goals, and 61 objectives.

*Assigning depth-of-knowledge level to mathematics test items*

The activity of assigning a depth-of-knowledge level to each objective was followed by the activity of assigning a depth-of-knowledge level to each of the mathematics test items. The teachers were asked to assign individually a depth-of-knowledge level to each of the mathematics test items, with instruction on how to use the coding matrix provided to them for recording their observations. Where an item involved more than one depth-of-knowledge level, the teachers were instructed to assign the highest level to the item. After the teachers had completed the activity, they were asked to discuss the depth-of-knowledge levels that they assigned to the mathematics test items, with the aim of reaching consensus. A consensus among the teachers is important because an item can have only one depth-of-knowledge level assigned to it. Therefore, the teachers should agree on the most appropriate depth-of-knowledge level for each item. After the teachers reached consensus, they were presented with the original depth-of-knowledge levels that had been assigned to the mathematics test items by PISA. The teachers were asked to compare their list with the PISA list and evaluate the depth-of-knowledge levels assigned to a pair of corresponding items on the two lists.

*Matching depth-of-knowledge level of mathematics test items with BMCS objectives*

The teachers were instructed to match the depth-of-knowledge levels of the mathematics test items with the objectives of BMCS, based on the condition that a student's responses to the test items provided information about what the student knew or could do with respect to an objective (Webb, 1999). The teachers wrote each item's depth-of-knowledge level in each row of an objective corresponding

to the item's column of a coding matrix. Each objective that was matched to a depth-of-knowledge level of an item was called a hit. Multiple hits were allowed, and no limit on the number of hits for an item was set. This meant that an item could be matched to more than one objective. However, after discussing among themselves on their individual work, the teachers were able to reduce their differences in the number of hits for an item to a noticeable extent. In accordance with Webb's (1999) method, the hits were used to compute mean, frequency, and percentage to evaluate the alignment of test items to BMCS. All of the statistics were computed for each teacher, and the mean was computed for each alignment criterion.

### *Field administrations of the mathematics test*

First, the mathematics test was trialled in a school in Bhutan, with the view to improving the test before administering its final version. Thirty-six students participated in the trial test. Only the qualitative analysis was performed with the trial test data. The qualitative analysis of the trial test data focussed on the following characteristics of the test: the suitability of the writing time, the ambiguity of the test item wordings, and the adequacy of the answer space. In addition, students were invited to make post-test comments, particularly their feelings about the test. A complete protocol of the psychometric analysis of the test could not be applied in analysing the trial test data for want of an adequate sample size. For instance, a minimum sample size of five test candidates is recommended for every item to obtain the following information: item *p* values, item-test correlations, item discrimination indices, and item-differential functioning test statistics (Crocker & Algina, 2008; Nunnally, 1978). However, as the test items were adapted from PISA, information about *item* p *values* and item difficulty parameter was available for use (OECD, 2009b).

Second, the final version of the mathematics test was administered to 1,500 Grade 10 students enrolled in 60 schools across Bhutan. Prior consents were obtained from the schools concerned, and test schedules were provided to the schools before administering the test. The test was administered by a group of test administrators who were trained to administer the test. The training was aimed at standardizing the test administration procedures so that the influence of different test administrators or test environments on the test candidates was avoided or reduced (Evers, 2001).

### *Scoring the response papers*

Response papers were evaluated and scored by the researcher in line with the test scoring guide that accompanied the 42 PISA items (OECD, 2009b). This ensured consistency in the scores across the papers and avoided the need for marker training and evaluation of *inter-rater agreement*.

The test scores were entered into computers by a trained data entry team of four members, entering 50 papers per day. The data entry was done in the morning, followed by a data cleaning session in the afternoon, with the data cleaned by using descriptive statistics, such as score frequencies, range, and mean. The errors detected during data screening were corrected by referring to the original response papers. This mode of data entry helped in minimising the errors committed during

entering the data into computers, as evidenced by fewer errors after a first couple of days.

### *Test reliability*

Classical test score theory assumes that each test taker has a true score that would be obtained if there were no errors in measurement (Aiken & Growth-Marnat, 2006; Kaplan & Saccuzzo, 2001; Nunnally, 1978). Errors in measurement are systematic or random. Kaplan and Saccuzzo (2001) noted that systematic errors in measurement are less likely to misguide an investigator to making wrong inferences than the random errors. Classical test score theory assumes that measurement errors are random and commonly identifies unconducive test environment, test fatigue, demotivated test participants, test validity, and so on, as the possible sources of the errors. The reliability estimate for the mathematics test was calculated by using Cronbach's coefficient alpha (Aiken & Growth-Marnat, 2006; Kaplan & Saccuzzo, 2001). Coefficient alpha was preferred over other statistic (e.g., test-retest, parallel forms, split-half, KR20, KR21) because it is "the most general method of finding estimates of reliability through internal consistency" (Kaplan & Saccuzzo, 2001, p. 113). In addition, the threats of systematic errors to the test data were minimised by following standard psychometric procedures in developing the test and by training test administrators on test administration modalities.



### *Calibrating test items*

The mathematics test items were calibrated by using the maximum likelihood estimation procedure available in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). Adams and Wilson (1996) presented an excellent discussion on the use of maximum likelihood estimation procedure in ConQuest, and the same is not discussed here. Among the different item response models that ConQuest is capable of fitting (Adams & Wilson, 1996; Wu et al., 2007), the partial credit model (Masters, 1982; Masters & Wright, 1997) was used to calibrate the test items because of its close fit with the types of items used in the mathematics test. The result from the item calibration was used in deciding whether an item was worth including in generating students' proficiency scores.

In addition to the sound psychometric properties of the test items, students' motivation during the test and sufficiency of the test writing time also affect the test validity. The missing response data were analysed to confirm or rule out the possibility of the influence of these factors on students' responses to the test items.

### *Estimating students' proficiency scores*

Students' proficiency scores were estimated by using *plausible values* generated with the plausible value estimation option available in ConQuest, with a mean of zero and a standard deviation of one. A succinct discussion on the need to use *plausible values* in large-scale assessments have been presented by Wu (2005), Mislevy, Beaton, Kaplan, and Sheenan (1992), and the OECD (2005c). For ease of interpretation, and to facilitate comparison with the performance of the OECD countries, the proficiency scores were transformed to the PISA scale, with a mean of 500 score points and a standard deviation of 100 units.

***Linking mathematics test to PISA 2003***

The performance scores of Grade 10 Bhutanese students on the mathematics test were linked to the PISA scale by using the *scaling constants* derived from the *common items* in the mathematics test and the PISA 2003 Mathematics Literacy test. Scaling constants were derived based on the approach of Kolen and Brennan (2004). Further, the OECD (2005c) presented a thorough technical discussion on linking different cycles of PISA by using the scaling constants derived from the common items. The link makes it possible to compare the performance of the Grade 10 Bhutanese students with the performance of the students across the countries who participated in the PISA 2003 mathematics literacy test. It is beyond the scope of the article to describe fully the procedures involved in linking different tests, and thus only those aspects relevant to the article are described here.

Kolen and Brennan (2004) presented the following equations for relating person and item parameters on tests $J$ and $I$:

$$\theta_{Ji} = A\theta_{Ii} + B \tag{1}$$

$$b_{Ji} = Ab_{Ij} + B \tag{2}$$

In Equations (1) and (2), $\theta$ denotes the student ability, and $b$ denotes the item difficulty parameter. The constants $A$ and $B$ are the scaling constants. Kolen and Brennan (2004) presented the following equations for deriving scaling constants from a set of common items in tests $I$ and $J$:

$$A = \frac{\sigma(b_J)}{\sigma(b_I)} \tag{3}$$

$$B = \mu(b_J) - A\mu(b_I) \tag{4}$$

In Equations (3) and (4), $\sigma$ and $\mu$ are the standard deviation and the mean of the item difficulty parameters of the common items in tests $I$ and $J$. Depending on the values of the scaling constants in Equations (3) and (4), the performance scores of Grade 10 Bhutanese students on the mathematics test can be linked to the performance scores of the students who participated in the PISA 2003 Mathematical Literacy test by using Equations (1) and (2). The linked scores make it possible to benchmark the two groups of students and hence their education systems. In this article, the benchmarks were established in terms of (a) the mean performance scores and (b) the percentage of students in each PISA Mathematics Proficiency Level.

*Linking error*

In principle, it is desirable to have a perfect match between the parameters of the common items across tests, so that the difference between any two parameters of the common items is zero. However, the item parameters of the common items change across tests. The change in the item parameters of the common items across tests introduces a linking error (OECD, 2005b). Therefore, the linking error is the standard error of the difference in the item difficulty parameters of the common

items across tests. The OECD (2005b) shows a method of computing the linking error by dividing the standard deviation of the difference by the square root of the number of common items.

## Result

### *Test validity*

*Objectives and domains of BMCS and PML*

First, the objectives of BMCS and PML contained similar keywords and similar verbs in emphasizing the expectations of students' mathematical knowledge and skills, indicating that the scope of the mathematical knowledge and skills expected of Grade 10 Bhutanese students and the PISA 2003 participants was similar. Both BMCS and PML expect students to learn and demonstrate similar mathematical knowledge and skills at the end of their schooling. For example, BMCS expects Bhutanese students to reason, communicate, and confidently use their mathematical knowledge and skills as they solve, describe, explore, and discover mathematical problems in various situations (Curriculum and Professional Support Division [CAPSD] & Bhutan Board of Examinations [BBE], 2007). Similarly, PML expects 15-year-olds to "analyse, reason, and communicate ideas effectively as they pose, solve, and interpret mathematical problems in a variety of situations" (OECD, 2004b, p. 24).

Second, the domains of BMCS and PML matched well, and some domains overlapped each other as shown in Table 1. For instance, *Numeration* in BMCS emphasises the following mathematical knowledge and skills: an understanding of number meanings, ordering and representing real numbers, and applying a variety of number theory concepts in solving problems. *Quantity* in PML emphasises the following mathematical knowledge and skills: an understanding of relative size, the recognition of numerical patterns, the use of numbers to represent quantities and quantifiable attributes of real-world objects, and estimation (OECD, 2004b). As mathematical knowledge and skills emphasized in *Numeration* and *Quantity* are similar, they are mapped together. Table 1 leads to the inference that the domains of BMCS and the domains of PML are similar in mathematical knowledge and skills.



Table 1. Match between BMCS and PML.

| | PISA Mathematics Domains | | | |
|---|---|---|---|---|
| Grade 10 Mathematics Curriculum Domain | Space and Shape | Change and Relationship | Quantity | Uncertainty |
| Numeration | | | √ | |
| Operation | | | √ | |
| Patterns | | √ | | |
| Measurement | √ | √ | | |
| Geometry | √ | √ | | |
| Data management and Probability | | √ | √ | √ |

Note: "√" indicates the match between the domains of BMCS and the PLM.

*Alignment of mathematics test and BMCS*

Based on alignment criteria mentioned earlier, the mathematics test and BMCS passed the categorical concurrence on four standards (Serial [Sl.] No. 2, 3, 4, and 6), the depth-of-knowledge consistency on four standards (Sl. No. 1, 2, 4, and 6), the range-of-knowledge on three standards (Sl. No. 2, 5, and 6), and the balance-of-representation on three standards (Sl. No. 3, 4, and 6). Table 2 presents the summary of the findings.

Overall, the mathematics test and the BMCS were partially aligned (50% to 67%), indicating that the test contained sufficient items to assess the mathematical knowledge and skills of Grade 10 Bhutanese students in line with what they were expected to know and do by BMCS.



***Field administration of the mathematics test***

The qualitative analyses of the students' response from the trial test data showed that the test was appropriately timed, that the test items were clearly worded, and that the response spaces for the items were sufficient.

For instance, few answer papers had predictable incorrect and/or missing responses to the test items located at the end of the test, indicating adequate test-writing time. Few students left the test room before the last five minutes of the test, confirming that the test-writing time was proportional to the mathematical knowledge and skills inherent in the test. The 2-hr writing time was also in close proximity to the average of 2-min writing time for a test item deduced by PISA (OECD, 2005c). Possible ambiguities in the item wordings of the test were identified by studying the clarifications sought by the students during the test session. Consequently, some errors were spotted in the diagrams and the graphs. No students had used a separate answer sheet for any question, and there was no spill over writing or writing in reduced font size in any answer space, indicating the adequacy of the answer space provided a priori in the test paper.

The errors observed from the trial test data were corrected in the final version of the test.

Table 2. Alignment of mathematics test and BMCS.

| SI. No. | Standards | Categorical Concurrence | Depth-of-Knowledge Consistency | Range of Knowledge | Balance of Representation |
|---|---|---|---|---|---|
| 1 | 10-Strand A-Number | No | Yes | No | No |
| 2 | 10-Strand B-Operations | Yes | Yes | Yes | No |
| 3 | 10-Strand C-Pattern | Yes | Weak | No | Yes |
| 4 | 10 -Strand D-Measurement | Yes | Yes | No | Yes |
| 5 | 10-Strand E-Geometry | No | Weak | Yes | Weak |
| 6 | 10-Strand F-Data Management and Probability | Yes | Yes | Yes | Yes |
| Percentage (%) | | 67 | 67 | 50 | 50 |

Note: *Yes* indicates alignment, *No* indicates non-alignment, and *weak* indicates marginal alignment.

### The test reliability

As reported earlier, the mathematics test used coefficient alpha as a measure of its reliability. The coefficient alpha for the mathematics test was 0.78, with a standard error of 0.01. This reliability estimate was deemed adequate for the intended purposes of the test, as observed by Evers (2001). Evers noted that a reliability coefficient between 0.70 and 0.80 is sufficient for making decisions on students' learning, and a reliability coefficient greater than 0.70 is good for research at group level; indicating the adequacy of the reliability coefficient of 0.78 of the test. Similar values have been obtained in PISA (OECD, 2009a) and the Third International Mathematics and Science Study (TIMSS) (International Association for the Evaluation of Educational Achievement, 2008).



### Calibrating test items

Item calibration showed that all items fitted well with the item response model except for one item with item-rest correlation of $-0.10$, indicating that the item functioned differently from the rest of the items in the test (Crocker & Algina, 2008). On close inspection of the mathematics test, the item had some typographical errors in its response options.

The analysis of the missing values in the test data by using SPSS (SPSS Inc, 2004) Missing Value Analysis (MVA) option showed that the data were missing completely at random (MCAR) as indicated by Little's MCAR test, $\chi^2 = 39103.219$, $df = 39039$, $p = 0.409$. The *Missing Patterns* table generated during MVA did not show any discernible response patterns. For instance, a not-reached item was expected to have an item immediately preceding it and the remaining items following it left unanswered, which was not visible in the *Missing Patterns* table. Two inferences from the observations were that the students remained motivated when writing the test and that the missing values in the data be treated as incorrect responses.

Overall, the mathematics test was reduced to 41 items from its initial number of 42 items.

### Linking mathematics test to PISA 2003

Table 3 presents the item parameters of the common items used in the mathematics test and the PISA 2003 Mathematical Literacy test. Using Equation (1), students' scores on the mathematics test were used in predicting their scores on the PISA 2003 mathematics test as follows:

$$P_{03} = \alpha(M_T) + \beta \tag{5}$$

In Equation (5), $P_{03}$ denotes the predicted scores on the PISA 2003 mathematics test for Grade 10 Bhutanese students, $M_{\mathrm{TS}}$ denotes the scores of Grade 10 Bhutanese students on the mathematics test, and $\alpha$ and $\beta$ are the scaling constants. Similar equations were used by Johnson and Owen (1998) for linking the National Assessment of Educational Progress (NAEP) and the TIMSS results.

Based on Equations (3) and (4), the scaling constants were derived by using the means and the standard deviations of the 11 common items shown in Table 3 as follows:

Table 3. Common items and their parameters as calibrated in the two tests.

| | PISA 2003 Mathematics Test | | Mathematics Test | | | |
|---|---|---|---|---|---|---|
| Item ID | Difficulty Estimate | Centred Difficulty Estimate | Difficulty Estimate | Centred Difficulty Estimate | Centred Difficulty Difference | Difference Squared |
| Q1(a) | −0.867 | −0.125 | −0.924 | −0.21 | −0.085 | 0.007225 |
| Q1(b) | −0.453 | 0.289 | −0.265 | 0.449 | 0.16 | 0.0256 |
| Q7(a) | −0.861 | −0.119 | −0.706 | 0.008 | 0.127 | 0.016129 |
| Q7(b) | −2.037 | −1.295 | −2.126 | −1.412 | −0.117 | 0.013689 |
| Q8(a) | 0.101 | 0.843 | 0.134 | 0.848 | 0.005 | 0.000025 |
| Q14(a) | −0.824 | −0.082 | −0.908 | −0.194 | −0.112 | 0.012544 |
| Q15(b) | 1.119 | 1.861 | 1.248 | 1.962 | 0.101 | 0.010201 |
| Q16(a) | −1.833 | −1.091 | −1.511 | −0.797 | 0.294 | 0.086436 |
| Q16(b) | −1.408 | −0.666 | −1.200 | −0.486 | 0.18 | 0.0324 |
| Q16(c) | 0.474 | 1.216 | 0.292 | 1.006 | −0.21 | 0.0441 |
| Q17(a) | −1.567 | −0.825 | −1.886 | −1.172 | −0.347 | 0.120409 |
| *M* | −0.742 | | −0.714 | | −0.00036 | |
| *SD* | 0.988 | | 1.005 | | 0.19203 | |

$$\alpha = \frac{\sigma_{P_{03}}}{\sigma_{M_T}} = 0.983 \tag{6}$$

$$\beta = \mu_{P_{03}} - \alpha(\mu_{M_T}) = -0.040 \tag{7}$$

In Equations (6) and (7), $\mu_{P_{03}}$, $\sigma_{P_{03}}$, $\mu_{M_T}$ and $\sigma_{M_T}$ denote the means and standard deviations of difficulty estimates of the 11 common items as used in the PISA 203 Mathematical Literacy test and mathematics test, respectively. Substituting the values of the scaling constants in Equation (5) yielded the following equation:

$$P_{03} = 0.983(M_{TS}) - 0.040 \tag{8}$$

Equation (8) is in logit metric with a mean of zero and a standard deviation of one. Because the PISA 2003 mathematics scores were standardised with a mean of 500 scores and a standard deviation of 100 units, Equation (8) was standardised as shown in Equation (9). Standardization of Grade 10 mathematics test scores with a mean of 500 and a standard deviation of 100 made the test scores comparable to the PISA 2003 Mathematical Literacy test scores.

$$P_{03} = (0.983(M_T) - 0.040) \times 100 + 500 \tag{9}$$

Once the scores on the mathematics test and the PISA 2003 are comparable, students' mathematical knowledge and skills can be interpreted in terms of the *PISA Proficiency Scale* described in the OECD (2005c, 2009a). The benefit of using the PISA Proficiency Scale to interpret students' mathematical knowledge and skills is that the performance of the students who participated in the mathematics test can be benchmarked with the performance of the students who participated in the PISA 2003 mathematical literacy test. This is the focus of this article.

### *Linking error*

The standard deviation of the difference in the item parameters of the 11 common items is equal to 0.192. Therefore, the linking error is 0.059 logit units. On the mathematics test scale with a mean of 500 and a standard deviation of 100, this linking error corresponds to 5.79. The linking error leads to overestimation of the mean of the test scores. Elaborate discussions on the properties of the linking error are presented in Kolen and Brennan (2004) and the OECD (2005b). In line with its properties, the linking error has been used for computing the variance of the sample statistic.

### *Benchmarks*

As described earlier, benchmarks are established by estimating the predicted mean score of Grade 10 Bhutanese students on the PISA 2003 Mathematical Literacy test. Table 4 presents the predicted mean score of Grade 10 Bhutanese students on the PISA 2003 Mathematical Literacy test, together with the mean scores of the countries that participated in PISA 2003. The data for the countries that participated in PISA 2003 are obtained from OECD (2004a, p. 358). The predicted mean performance score of Grade 10 Bhutanese students on the PISA 2003 Mathematical Literacy test is 361 ($SE = 4.1$). The mathematical knowledge and skills of Grade 10 Bhutanese students are comparable with those of the students across the countries that participated in PISA 2003. For instance, Grade 10 Bhutanese students' mean score is greater than the mean scores of Indonesia (360), Tunisia (359), and Brazil (356). However, a word of caution has to be emphasized while comparing the scores from two different tests. As emphasized authoritatively by Johnson and Owen (1998), comparisons of scores from two linked tests should be viewed as estimates and not as substituting one test for the other.

Notwithstanding the cautionary note, such estimates provide guidelines for cross-country fertilization of innovative technologies in school education. For example, countries with similar characteristics can learn from their weaknesses and strengths through exchange of knowledge and skills related to school effectiveness.

To provide further insight into students' proficiency levels, the percentage of students at each proficiency level is calculated. Table 5 presents the result of the computation.

Table 5 shows that one fourth of the students scored below Level 1 of the PISA proficiency scale. The OECD considers the students scoring below Level 1 of the PISA proficiency scale as being at risk of not achieving the mathematical knowledge and skills that will enable them to participate fully in society beyond school (Thomson & De Bortoli, 2008).

On the whole, Table 5 shows that the majority of students performed around proficiency Levels 1 and 2, as depicted by a fewer number of students who managed to attain higher levels of the proficiency scale. This pattern is consistent with most of the OECD countries as shown in Figure 1.

Figure 1 shows the percentage of students at each level of the PISA mathematics proficiency scale across the countries that participated in PISA 2003. The chart is constructed with data from PISA 2003 (OECD, 2004a, p. 354). In addition, shown in the same chart is the percentage of Grade 10 Bhutanese students based on their predicted scores on the PISA 2003 Mathematical literacy test. On average, only about a third of students across the OECD countries attained Levels 5 and 6 on the PISA Proficiency Scale (OECD, 2005a).

Table 4. Mean mathematics scores of Bhutan and PISA 2003 countries.

| Country | *M* | *SE* |
|---|---|---|
| Hong Kong-China | 550 | 4.5 |
| Finland | 544 | 1.9 |
| Korea | 542 | 3.2 |
| Netherlands | 538 | 3.1 |
| Liechtenstein | 536 | 4.1 |
| Japan | 534 | 4 |
| Canada | 532 | 1.8 |
| Belgium | 529 | 2.3 |
| Switzerland | 527 | 3.4 |
| Macao-China | 527 | 2.9 |
| Australia | 524 | 2.1 |
| New Zealand | 523 | 2.3 |
| Czech Republic | 516 | 3.5 |
| Iceland | 515 | 1.4 |
| Denmark | 514 | 2.7 |
| France | 511 | 2.5 |
| Sweden | 509 | 2.6 |
| Austria | 506 | 3.3 |
| Germany | 503 | 3.3 |
| Ireland | 503 | 2.4 |
| OECD average | 500 | 0.6 |
| Slovak Republic | 498 | 3.3 |
| Norway | 495 | 2.4 |
| Luxembourg | 493 | 1 |
| Hungary | 490 | 2.8 |
| Poland | 490 | 2.5 |
| OECD total | 489 | 1.1 |
| Spain | 485 | 2.4 |
| United States | 483 | 2.9 |
| Latvia | 483 | 3.7 |
| Russian Federation | 468 | 4.2 |
| Italy | 466 | 3.1 |
| Portugal | 466 | 3.4 |
| Greece | 445 | 3.9 |
| Serbia | 437 | 3.8 |
| Turkey | 423 | 6.7 |
| Uruguay | 422 | 3.3 |
| Thailand | 417 | 3.0 |
| Mexico | 385 | 3.6 |
| BHUTAN | 361 | 4.1 |
| Indonesia | 360 | 3.9 |
| Tunisia | 359 | * |
| Brazil | 356 | 4.8 |

Note: *not available. Data for countries other than Bhutan are obtained from OECD (2004a, p. 358).



## Discussion and conclusion

The findings and the methodological approach have key implications for benchmarking student performance as a basis for key interventions at the level of a country's education system.

The aim of this article was to demonstrate a method of benchmarking a country's education system with the education systems of the countries that

Table 5. Percentage of students at each proficiency level.

| Proficiency Level | Percentage | *SE* |
|---|---|---|
| Below 1 | 27.03 | 2.03 |
| 1 | 35.16 | 1.67 |
| 2 | 26.62 | 1.44 |
| 3 | 8.092 | 1.26 |
| 4 | 2.7 | 0.65 |
| 5 | 0.30 | 0.14 |
| 6 | 0.10 | 0.08 |

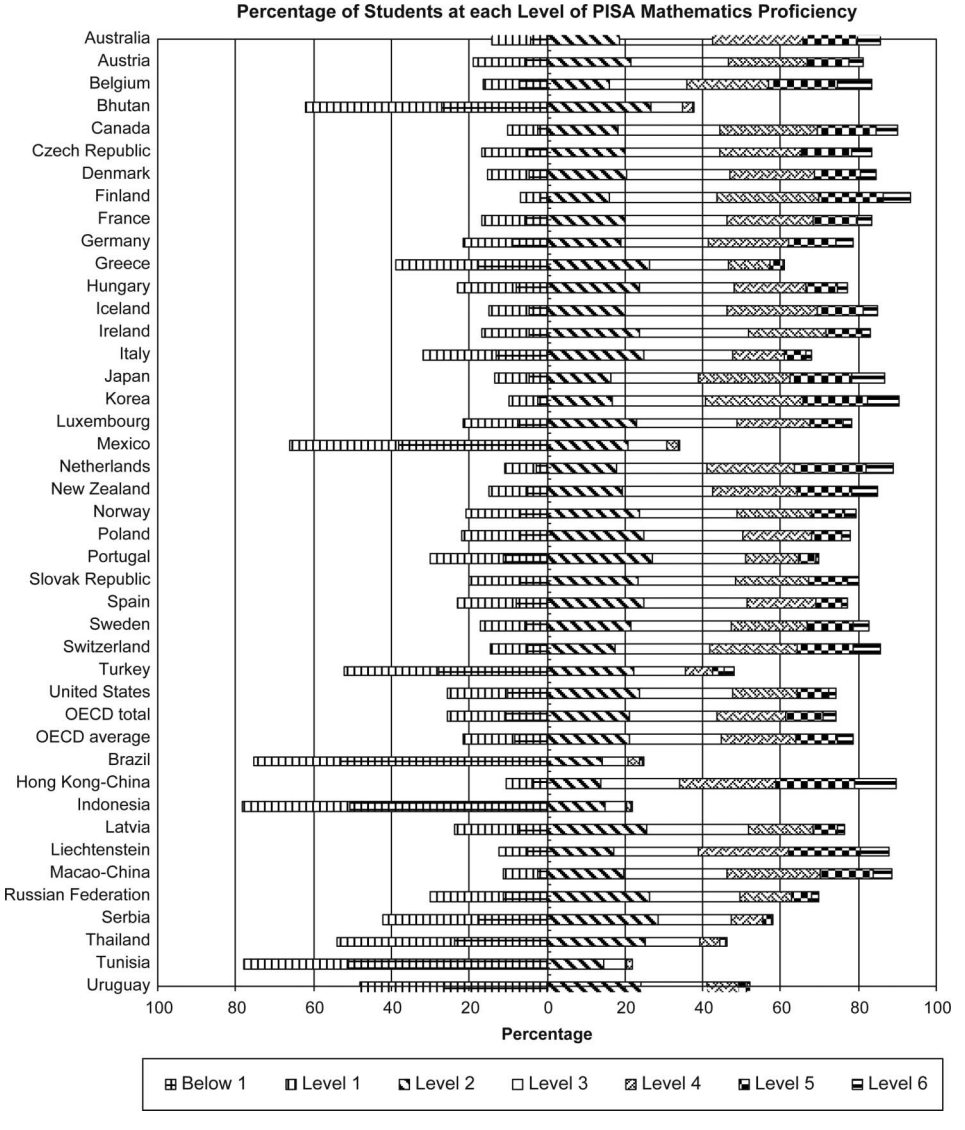


Figure 1. Percentage of students at proficiency levels.
Note: Data for countries other than Bhutan are obtained from OECD (2004a, p. 354).

participated in PISA 2003. It has been demonstrated that it is plausible to benchmark a country's education system by constructing a test from the PISA-released items. A test constructed with the PISA-released items has to be valid and reliable. This article has demonstrated two methods of ensuring that the mathematics test constructed from the PISA-released items is valid. The following two approaches have been used: (a) comparing PML to BMCS and (b) evaluating the alignment of the mathematics test items to the BMCS. These two methods showed that PML and BMCS had similar goals, objectives, and content scope; and that mathematics test items achieved partial alignment to the BMCS. The similarities between PML and BMCS and the partial alignment of the mathematics test items to the BMCS effectively indicate the validity of the mathematics test.

The article presented a method to link the mathematics test to the PISA 2003. It has been shown that the scores on the mathematics test can be used to predict corresponding scores on the PISA 2003. Based on the predicted scores of the students on the PISA 2003, benchmarks have been established in terms of the mean score of the students and the percentage of the students in the PISA Proficiency Level. The benchmark in terms of the mean score has the potential to help a country to identify other countries with similar performance. Such information is helpful in assisting countries in identifying prospective partner countries for collaboration in developing school improvement programmes and research. The benchmark in terms of the PISA Proficiency Levels can offer a country an insight into its students' knowledge and skills of a school subject and a broad overview of knowledge and skills of the students in other countries. Such insight can help a country in developing teaching and learning strategies that help students to acquire critical thinking skills that underpin the PISA Proficiency Levels. Kyriakides (2006, p. 490) refers to such benchmarks as "cross-national perspective". These benchmarks have been empirically demonstrated in this study.

The benefits of using PISA-released items to benchmark a country's education system with the education systems of other countries that participated in PISA cycles are substantial. However, the benefits come at a price of precision, as is the case with any prediction. The potential sources of error are linking error and alignment of test items to school curriculum standards. Where it is feasible for a country to participate in PISA, it would be interesting to conduct a study by using the method described in this article. The performance scores of the students on PISA can be compared with their predicted scores on PISA that are derived from a test developed from a set of the PISA-released items. Such a study will either validate or invalidate the findings discussed in this article.



## Notes on contributors

Gembo Tshering (gtshering@students.latrobe.edu.au) is a PhD student at the Faculty of Education of La Trobe University in Australia. His PhD study relates to school effectiveness, large-scale assessments, international comparisons, school examinations, and educational policy.

Vaughan Prain (v.prain@latrobe.edu.au) is a Professor at the Faculty of Education of La Trobe University in Australia. His research interests include the role of representational construction in learning in secondary school science and the use of digital technologies in secondary education.

## References


Adams, R.J., & Wilson, M. (1996). Formulating the Rasch Model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Measurement: Theory into practice* (Vol. 3, pp. 143–166). Norwood, NJ: Ablex Publishing Corporation.

Aiken, L.R., & Growth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Boston, MA: Pearson Education Group.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational research Association.

Creemers, B.P.M. (2006). The importance and perspectives of international studies in educational effectiveness. *Educational Research and Evaluation*, *12*, 499–511.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York, NY: Rinehart and Winston.

Curriculum and Professional Support Division (CAPSD) & Bhutan Board of Examinations (BBE). (2007). *Syllabus for 9 & 10: Bhutan Certificate for Secondary Education*. Retrieved from http://www.education.gov.bt/Secretariat/syllabus.htm

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Evers, A. (2001). The Revised Dutch Rating System for Test Quality. *International Journal of Testing*, *1*, 155–182.

Gronlund, N.E., & Waugh, C.K. (2009). *Assessment of student achievement* (9th ed.). Upper Saddle River, NJ: Pearson Education.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

International Association for the Evaluation of Educational Achievement. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Centre, Lynch School of Education, Boston College.

Johnson, E.G., & Owen, E. (1998). *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A technical report* (No. NCES 98–499). Washington, DC: National Center for Educational Statistics.

Johnson, E.G., & Phillips, G.W. (1998). *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eight-grade results* (No. 98–500). Washington, DC: National Center for Educational Statistics.

Kaplan, R.M., & Saccuzzo, D.P. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

Kyriakides, L. (2006). Introduction. International studies on educational effectiveness. *Educational Research and Evaluation*, *12*, 489–497.

Linn, R.L. (2002). Validation of the uses and interpretations of results of state assessment and accountability systems. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 27–48). London, UK: Lawrence Erlbaum Associates.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Masters, G.N., & Wright, B.D. (1997). The partial credit model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer.

McGaw, B. (2008). Further reflections. *Assessment in Education: Principles, Policy & Practice*, *15*, 279–282.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

Nardi, E. (2008). Cultural biases: A non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice*, *15*, 259–266.

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
Organisation for Economic Co-operation and Development. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Retrieved from http://lysander.sourceoecd.org/vl=6357152/cl=16/nw=1/rpsv/ij/oecdthemes/99980029/v1999n1/s1/p1
Organisation for Economic Co-operation and Development. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2002). *PISA 2000 technical report*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2004a). *Learning for tomorrow's world: First result from PISA 2003*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2004b). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2005a). *Are students ready for a technology-rich world? What PISA studies tell us*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2005b). *PISA 2003 data analysis manual: SPSS*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2005c). *PISA 2003 technical report*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Retrieved from http://www.oecd.org/pages/0,3417,en_32252351_32236191_1_1_1_1_1,00.html
Organisation for Economic Co-operation and Development. (2007). *PISA 2006: Science competencies for tomorrow's world: Volume 1 analysis*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2009a). *PISA 2006 technical report*. Paris, France: Author.
Organisation for Economic Co-operation and Development. (2009b). *Take the test: Sample questions from OECD's PISA assessments*. Paris, France: Author. Retrieved from www.sourceoecd.org/education/9789264050808
Osta, I. (2007). Developing and piloting a framework for studying the alignment of mathematics examinations with the curriculum: The case of Lebanon. *Educational Research and Evaluation*, *13*, 171–198.
Phillips, G.W. (2007). *Expressing international educational achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS*. Washington, DC: American Institute for Research.
Prais, S.J. (2003). Cautions on OECD'S recent educational survey (PISA). *Oxford Review of Education*, *29*, 139–163.
Ross, K.N. (1992). Sample design for international studies of educational achievement. *Prospects*, *22*, 305–316.
SPSS Inc. (2004). SPSS for Windows (Version 18.0) [Computer software]. Chicago, IL: SPSS Inc.
Thomson, S., & De Bortoli, L. (2008). *Exploring scientific literacy: How Australia measures up: The PISA 2006 survey of students' scientific, reading and mathematical literacy skills*. Camberwell, Victoria, Australia: ACER. Retrieved from http://www.acer.edu.au/documents/PISA2006_Report.pdf
Van der Linden, W.J., & Hambleton, R.K. (Eds.). (1997). *Handbook of modern item response rheory*. New York, NY: Springer.
Wagemaker, H. (2008). Choices and trade-offs: Reply to McGaw. *Assessment in Education: Principles, Policy & Practice*, *15*, 267–278.
Webb, N.L. (1997). *Research Monograph No. 8: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.
Webb, N.L. (1999). *Research Monograph No. 18: Alignment of science and mathematics standards and assessments in four states*. Madison, WI: National Institute for Science Education.

Webb, N.L. (2006). Identifying content for student achievement tests. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S.A. (2007). *ACER ConQuest: Generalised item response modelling software* (Vol. 2). Melbourne, Australia: ACER Press.